CRM 86-214 / October 1986

RESEARCH MEMORANDUM

FROM HANDS-ON MEASUREMENT TO JOB PERFORMANCE: THE ISSUE OF GENERALIZABILITY

Paul W. Mayberry

A Division of CNA Hudson Institute

CENTER·FOR·NAVAL·ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

Work conducted under contract N00014-83-C-0725.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS AD-A178551
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) CRM 86-214	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Center for Naval Analyses 6b. OFFICE SYMBOL (If applicable) CNA	7a. NAME OF MONITORING ORGANIZATION Commandant of the Marine Corps (Code RDS)
6c. ADDRESS (City, State, and ZIP Code) 4401 Ford Avenue Alexandria, Virginia 22302-0268	7b. ADDRESS (City, State, and ZIP Code) Headquarters, Marine Corps Washington, D.C. 20380
8a. NAME OF FUNDING / ORGANIZATION Office of Naval Research 8b OFFICE SYMBOL (If applicable) ONR	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-83-C-0725
8c. ADDRESS (City, State, and ZIP Code) 800 North Quincy Street Arlington, Virginia 22217	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 65153M; PROJECT NO C0031; TASK NO; WORK UNIT ACCESSION NO
11. TITLE (Include Security Classification)	
12. PERSONAL AUTHOR(S)	
13a. TYPE OF REPORT Final 13b. TIME COVERED FROM TO	14. DATE OF REPORT (Year, Month, Day) October 1986 15. PAGE COUNT 64
16 SUPPLEMENTARY NOTATION	

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	09	
12	01	

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Analysis of variance, Aptitude tests, Enlisted personnel, Estimates, G-Theory (Generalizability Theory), Hands-on tests, Job training, Marine Corps personnel, Mathematical analysis, MOS (military occupational specialty), Performance (human), Performance tests, Qualifications, Reliability, Statistical analysis, Specialties

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

The generalization from hands-on test scores to performance in a military occupational specialty is threatened by many potential sources of error within the measurement process. Such sources of error can include scoring inconsistencies by test administrators, testing over a long period, and diverse test content. This analysis estimates the influence of these factors on the hands-on scores for three Marine Corps specialties. Estimates of test reliability are discussed in light of the effect of the measurement factors on the hands-on scores. Research designs to assess specific issues of reliability are proposed for the full-scale administration of hands-on tests to the Infantry occupational field.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Lt.Col. G. W. Russell	22b TELEPHONE (Include Area Code) (202) 694-3491 22c OFFICE SYMBOL RDS-40

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted

All other editions are obsolete

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

A Division of CNA Hudson Institute

CENTER·FOR·NAVAL·ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

10 November 1986

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 86-214

Encl: (1) CNA Research Memorandum 86-214, "From Hands-On Measurement to Job Performance: The Issue of Generalizability," by Paul Mayberry, October 1986

1. Enclosure (1) is forwarded as a matter of possible interest.

2. The Job Performance Measurement Project is a joint-service effort to develop good measures of job performance that will be used as part of the process to establish enlistment standards. Hands-on tests are generally considered to be the most definitive measures of performance. However, the generalization from hands-on test scores to performance in a military specialty can be threatened by many potential sources of error. This Research Memorandum identifies many of these factors and estimates their impact on hands-on test scores for three Marine Corps specialties. Specific attention is given to the resulting implications for the administration of hands-on tests to the Infantry occupational field.

William H. Sims
Director, Manpower and Training Program
Marine Corps Operations
Analysis Group

Distribution List:
Reverse Page

Subj: Center for Naval Analyses Research Memorandum 86-214

Distribution List

SNDL

A1	ASSTSECNAV MRA
A1	DASN - MANPOWER (2 copies)
A6	HQMC MPR
	Attn: Deputy Chief of Staff for Manpower (2 copies)
	Attn: Director, Personnel Procurement Division (2 copies)
	Attn: Director, Manpower Plans and Policy Division (2 copies)
	Attn: Director, Personnel Management Division (2 copies)
A6	HQMC TRNG (2 copies)
A6	HQMC RD&S (2 copies)
A6	HQMC RA (2 copies)
A6	HQMC AVN (2 copies)
E3D1	CNR
E3D5	NAVPERSRANDCEN
	Attn: Director, Manpower and Personnel Laboratory
	Attn: Technical Library
FF38	USNA
	Attn: Nimitz Library
FF42	NAVPGSCOL
FF44	NAVWARCOL
FJA1	COMNAVMILPERSCOM
FJB1	COMNAVCRUITCOM
FT1	CNET
V12	CG MCDEC

OPNAV

OP-01
OP-11
OP-12
OP-13

Other

Joint Service Job Performance Measurement Working Group (13 copies)

CRM 86-214 / October 1986

FROM HANDS-ON MEASUREMENT TO JOB PERFORMANCE: THE ISSUE OF GENERALIZABILITY

Paul W. Mayberry

Marine Corps Operations Analysis Group

A Division of CNA Hudson Institute

CENTER·FOR·NAVAL·ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

ABSTRACT

The generalization from hands-on test scores to performance in a military occupational specialty can be threatened by many potential sources of error within the measurement process. Such sources of error can include scoring inconsistencies by test administrators, testing over a long period, and diverse test content. This analysis estimates the influence of these factors on the hands-on scores for three Marine Corps specialties. Estimates of test reliability are discussed in light of the effect of the measurement factors on the hands-on scores. Research designs to assess specific issues of reliability are proposed for the full-scale administration of hands-on tests to the Infantry occupational field.

EXECUTIVE SUMMARY

BACKGROUND

Estimating how well individuals will perform their job responsibilities based on scores from hands-on job-performance tests is basically a function of test design and task selection. However, the inference process can be contaminated by potential shortcomings in the measurement process that are unrelated to test construction. To the extent that these measurement factors influence the hands-on test scores, the generalization from test performance to performance in the military occupational specialty (MOS) is weakened. Research findings based on inconsistent test scores are not generalizable because opposing results could be found under different circumstances, conditions, or occasions. Analyses that demonstrate the quality of measurement and establish the reliability of the hands-on tests are essential if such tests are to result in appropriate generalizations. Therefore, an effort should be made to identify any influential measurement factors and determine the magnitude of their impact on the hands-on scores.

An extension of classical reliability theory called Generalizability Theory (G-theory) addresses this concern. The conceptual framework of G-theory is based on the partitioning of observed score variance into as many components as the design of the study allows. In this manner, G-theory is capable of identifying specific aspects of the measurement process that give rise to the greatest degree of error. In addition, G-theory allows for the estimation of overall reliability (or generalizability) coefficients.

G-theory analyses were applied to data collected for a feasibility study of job-performance measurement conducted in 1981. Hands-on tests were developed for three Marine Corps MOSs: Ground Radio Repair, Automotive Mechanic, and Infantry Rifleman. Three measurement factors that might systematically contribute to the variance of hands-on scores were identified; although all three factors were not necessarily present in the research designs for the three MOSs. These measurement factors were the test administrator, testing occasion, and test content.

FINDINGS

Estimates of the mean squares and variance components for each MOS research design were computed. In general, the residual variance-component estimates (the unexplainable variance in the hands-on scores) were large. This is unfortunate because this is the term that should be kept to a minimum to enhance the reliability of the hands-on tests. Four different reliability estimates were computed for each MOS; these are presented in table I. Moderately high reliabilities were found for the Ground Radio Repair test and very low reliabilities found for the other two specialities.

These large residual variance-component estimates and low reliabilities imply that the research-design models were not correctly specified and therefore did not include the appropriate measurement factors to account for the significant residual variances. In the case of the three specialties examined in this study, the issue may not be so much with misspecification of the model as with the inappropriateness of the data-collection effort. The data were not specifically collected for G-theory analyses. The purpose of G-theory analysis is to be able to estimate as many variance components as possible and the 1981 data-collection designs did not allow for the estimation of many important variance components.

As a result, experimental designs that correct this problem are proposed for the testing of the Infantry occupational field, the first stage of the full-scale Marine Corps Job Performance Measurement Project. Because the use of experimental designs to assess reliability (i.e., G-theory analysis) is extremely difficult to coordinate, expensive and time consuming to administer, and often disruptive to the personnel, it is recommended that G-theory mini-studies be conducted on limited samples.

Although it was not a specific finding in this study, the inconsistency of the test administrators is considered to be a major threat to the fidelity of hands-on measurement of job performance. Explicit examination of the administrators' scoring strategies and practices is the central focus of the G-theory experimental designs. Three G-theory studies are proposed to ad-

TABLE I

ALTERNATIVE ESTIMATES OF RELIABILITY FOR THREE HANDS-ON TESTS

	Ground Radio Repair	Automotive Mechanic	Infantry Rifleman
Number of items	10 boards	6 tasks	12 tasks
Inter-item correlation			
Mean	0.28	0.12	0.14
Range	0.00 to 0.59	0.01 to 0.27	–0.15 to 0.64
Alpha estimate	0.80	0.40	0.65
ANOVA[a] estimate	0.79	- -[b]	0.56
G-theory estimate			
Relative	0.80	0.37	0.64
Absolute	0.77	0.25	0.58

a. Analysis of variance procedure.

b. The ANOVA reliability estimate for the Automotive Mechanic specialty could not be determined because the within-subject mean square was greater than the between-subjects mean square.

dress the question of administrator consistency. The first design addresses consistency between administrators. The mini-study also examines the extent to which alternate forms of the hands-on test are parallel, and whether taking one form of the test before the other results in an order effect. Two other designs that examine the questions of administrator consistency over time and administrator consistency across bases or testing locations are proposed.

CONCLUSIONS

- Generalizing from performance on the hands-on tests to performance in the three MOSs is limited, given the magnitude of the residual variance in the hands-on scores.
- Experimental designs that address the impact of specific measurement factors on the variance of the hands-on scores should be developed. Such designs allow for the simultaneous consideration of several factors and their interactions, as well as the calculation of generalizability coefficients.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF ILLUSTRATIONS

LIST OF TABLES

SECTION 1

INTRODUCTION

Estimating how well individuals will perform their job responsibilities based on scores from hands-on job-performance tests is basically a function of test design and task selection. However, the inference process can be contaminated by potential shortcomings in the measurement process that are unrelated to test construction. To the extent that these measurement factors influence the hands-on test scores, the generalization from test performance to performance in the military occupational specialty (MOS) is weakened. Research findings based on inconsistent test scores are not generalizable because opposing results could be found under different circumstances, conditions, or occasions. Analyses that demonstrate the quality of measurement and establish the reliability of the hands-on tests are essential if such tests are to result in appropriate generalizations. Therefore, an effort should be made to identify any influential measurement factors and determine the magnitude of their impact on the hands-on scores.

The analysis documented in this research memorandum estimates the influence of the measurement factors on the hands-on test scores. Estimates of reliability are examined in light of the effect of the measurement factors on the hands-on scores. Data from the Marine Corps Job Performance Measurement (JPM) feasibility study conducted for three selected MOSs are presented. Implications for the full-scale Marine Corps effort to develop and administer hands-on job performance tests to the Infantry occupational field are discussed.

RELIABLE PERFORMANCE MEASUREMENT

In contrast to paper-and-pencil testing, hands-on assessment can be affected by several potential sources of error within the measurement process. For example, a test of supposedly homogeneous content is often administered on different occasions and scored by a variety of raters. The

confounding of these factors - internal consistency, testing over time, and rater reliability - must be examined for their possible impact on an individual's observed test score.

In examining the potential effects of the measurement process on an individual's observed score, one must first look at the factors that compose this score. In the classical definition of measurement, an individual's observed score (X) is a function of two components: true score (T) and error (E), X = T + E. The reliability of an instrument is simply the squared correlation between true and observed scores. The larger the error component of the observed score, the larger the discrepancy between the true and observed scores and, accordingly, the smaller their correlation. In theory, the error component is assumed to be undifferentiated and univariate, although operationally it is known that errors can result from multiple sources. Treating error as undifferentiated can lead to unwarranted confidence in the dependability of the performance measure because the error may include a systematic component caused by certain aspects of the measurement process. Thus, it would be useful to identify potential factors of the measurement process that systematically contribute to the error component.

APPLICATION OF GENERALIZABILITY THEORY

An alternative to the classical reliability approach is Generalizability Theory (G-theory) [1]. G-theory begins with the recognition that all measurement is imperfect and proceeds to emphasize that errors result from multiple sources and are relative with respect to the total variance. The conceptual framework of G-theory is based on partitioning observed-score variance into as many components as the design of the study allows - in particular, more than the two sources purported in classical reliability theory. In this manner, G-theory is capable of pinpointing specific aspects of the measurement process that give rise to the greatest degree of error. Appropriate corrective actions can then be taken to minimize the further influence of these factors on the observed scores.

Estimating Reliability From Test Variance

The statistical underpinning for analyzing data for a G-theory study is the analysis of variance (ANOVA) procedure. The ANOVA partitions the variation of observed scores into separate components corresponding to main effects and their interactions. The factors (or facets) included in the analysis should represent meaningful elements within the measurement process that possibly contribute to the overall variation of hands-on scores.

In the simplest case, reliability of a hands-on test can be analyzed within the ANOVA framework by using two factors: subjects and items. The source of variation among the observed scores is a function of both between- and within-subject differences. Between-subject variance reflects systematic individual differences and also includes some average error of measurement for each subject. Within-subject variance estimates the error of measurement term and includes variation caused by both differences in the difficulty or content of items and the interaction of the subject and item factors.

Reliability ($\rho_{XX'}$) is expressed as the ratio of true-score variance to total observed-score variance (true-score variance plus error variance):

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \quad . \tag{1}$$

In the ANOVA model, these variances are estimated by the mean squares (MS) of the between- and within-subject factors:

Source of variation	MS	E(MS)
Between subjects	MS between subjects	$\sigma_T^2 + \sigma_E^2$
Within subject	MS within subject	σ_E^2

Substituting the mean squares from the above table for the variances of equation 1 and simplifying the equation results in the standard ANOVA

formula for reliability:

$$\rho_{XX'} = 1 - \frac{MS\ within\ subject}{MS\ between\ subjects} \quad . \tag{2}$$

Estimating Reliability From a G-Theory Perspective

Within the context of G-theory, the estimation of a reliability (or generalizability) coefficient is a secondary concern compared to the information resulting from the relative evaluation of the variance components. Summary indices that are analogous to the classical theory estimates do exist, but they vary depending on the types of decisions to be made from the data and the specification of the measurement model.

Decisions concerning individuals' performance are often made by normative comparisons of performance. These relative decisions require confidence in the ability of the measurement scale to maintain the generalizability of individual's *rank ordering* across all factors involved in the measurement process. For example, test-administrator inconsistencies in scoring performance across individuals will certainly alter the rank ordering of subjects. Therefore, this factor (administrator-by-subject) should be reflected in any relative generalizability estimate. These relative types of decisions are contrasted to decisions that are not dependent on normative performance levels but rather rely on comparisons against some *explicit performance standard.* The focus is now on the absolute level of performance itself and not performance relative to other individuals. In this case, mean differences across administrators as well as their inconsistencies in ranking individuals should be considered as error. It follows that the estimation of the generalizability coefficients should reflect the practical differences inherent in the specific utilization of the research findings.

The definition of the error variance for relative or absolute decisions is dependent on the specific design of the G-study. The generalizability coefficient is simply the proportion of the observed-score variance that is

attributable to systematic individual differences. The magnitude of this proportion changes as the variance components included in the error term are refined or expanded. The generic computational form for relative decisions is as follows:

$$\rho^2(Rel) = \frac{\sigma^2_{Subject}}{\sigma^2_{Subject} + \Sigma\sigma^2_{Subject\ interaction\ terms/n'}} \tag{3}$$

where

$n' =$ the respective sampling frequency of each source of error.

For the basic subject-by-item design, this reliability estimate is identical to KR-20 for dichotomously scored items and Cronbach's alpha for other metrics [2]. The absolute generalizability coefficient is defined as:

$$\rho^2(Abs) = \frac{\sigma^2_{Subject}}{\sigma^2_{Subject} + \Sigma\sigma^2_{Main\ effects\ and\ interaction\ terms/n'}} \tag{4}$$

where

$n' =$ the respective sampling frequency of each source of error.

As stated earlier, this estimate is somewhat more conservative than the relative estimate as evidenced by the inclusion of more variance components in the error term.

SECTION 2

PROCEDURES

In 1981, the Marine Corps conducted a study to evaluate the feasibility of validating enlistment standards against job performance [3]. Hands-on tests of job performance were developed for three military occupational specialities (MOSs):

- Ground Radio Repair: high technical requirements, 37 weeks of formal school training
- Automotive Mechanic: moderate technical requirements, 13 weeks of formal school training
- Infantry Rifleman: low technical requirements, 5 weeks of formal school training.

Only first-term enlistments were included in the study. The test administrators for each specialty were senior Marine Corps enlisted personnel with relevant job experience in their respective fields. The testing for all three specialties was conducted over a 3-month period.

The results are presented separately for the three occupational specialties because the design of the G-study is slightly different for each. For all three specialties, the specific measurement factors included in the analysis are briefly discussed. The results of the ANOVA and the variance-component estimates are presented and discussed in light of the various reliability estimates that arise from each.

SPECIFICATION OF THE GENERAL LINEAR MODEL

The factors included in the general linear model should represent meaningful elements within the measurement process that are thought to contribute to the overall variation of the performance scores. Three factors were identified that might systematically contribute to the variance of hands-on scores, although all three were not necessarily present in each of the three MOSs. These factors were test administrator, testing occasion, and test content.

The administrator factor is included in the analysis to account for differences among the raters in consistently assigning performance scores. Despite efforts to train administrators to rate performance in a consistent manner and to an appropriate standard, variations across administrators are likely to occur. This factor represents the effects of administrator miscalibration and drift.

Testing occasion represents a time dimension, reflecting when an individual was administered the hands-on test. Occasion was thought to contribute to the variance in test scores because of the possible compromise of test content to subjects tested late in the testing period. In addition to this test-security concern, later subjects may also have an advantage because of more on-the-job experience that may be directly relevant to the test content. The occasion variable was created by equally dividing the sample into first and second testing occasions according to testing date.

The test-content factor is included to partial out the variance specifically attributable to the heterogeneity of test content. To the extent that the subtests are not of equal difficulty or do not measure the same construct, variance in the observed scores attributable to these effects will be large.

The interactions of these factors also control for potential sources of systematic error. The interaction of the occasion and the administrator factors

represents administrator drift over time. The interaction of test content and administrator reflects administrators' inability to consistently rate performance across different content areas. The interaction of test content and occasion is indicative of differential performance on content areas over time.

A cautionary note must be given before the general linear models are described for each of the MOSs. The G-theory analyses were designed after the data for the feasibility study had been collected. In this regard, the designs tend to be unbalanced, nested, and contain few within-subject factors. The requisites for the data-collection stage for estimating all variance components of the G-theory analyses are discussed in the section 4. In addition, the factors included in the general linear models were considered to be random. Generally, subjects were representative of their respective MOS, although not strictly randomly selected. Test administrators also tended to be representative of the MOS job experts but not necessarily randomly selected. There was no reason to assume that subjects were in any way systematically assigned to the various levels of each factor, that is, test administrator or testing occasion. It was intended that inferences drawn from these research findings would extend to the populations of MOS-specific test administrators, testing occasions, and test content.

Ground Radio Repair

The hands-on test for the Ground Radio Repair specialty consisted of troubleshooting ten circuit boards from a novel piece of radio equipment. A total of 210 minutes, with up to 30 minutes for each board, was allowed. Some examinees were not able to work on all boards because of the total time limit. For each board, the examinee was instructed to identify the symptom (worth 2 points), faulty circuit (up to 4 points), and faulty component (up to 8 points). A total score for each individual was calculated as the sum of these three subscales. Examinees were encouraged to guess when they had narrowed the choice of circuits and components.

The Ground Radio Repair personnel were tested by one of five administrators. A testing occasion variable was created by dividing the sample

into equal groups according to their testing date. All administrators tested on both occasions. The design was unbalanced with respect to examinees within administrators and occasions:

		Administrator					
		1	2	3	4	5	Total
Occasion	1	7	5	10	10	12	44
	2	9	8	9	11	8	45
	Total	16	13	19	21	20	89

Examinees were nested within administrator and occasion, but administrator, occasion, and circuit boards were all crossed. The general linear model for the design was:

$$X = O + A + OA + S(OA) + B + OB + AB + OAB + BS(OA) + e \qquad (5)$$

where

X = observed score
O = occasion
A = administrator
OA = occasion-by-administrator interaction
$S(OA)$ = subjects within occasion and administrator
B = circuit board

OB = occasion-by-board interaction
AB = administrator-by-board interaction
OAB = occasion-by-administrator-by-board interaction
$BS(OA)$ = board-by-subject interaction
e = error.

Automotive Mechanic

The Automotive Mechanic hands-on test consisted of four major content areas. The content areas, tasks within these areas, and time limits were as follows:

- Major engine tuneup (120 minutes)
  - Compression
  - Coil
  - Vacuum
  - Precision timing
- Alternator output and battery (30 minutes)
  - Alternator
  - Battery
- Wheel and brake maintenance (60 minutes)
- Equipment repair order (completed as part of other tasks)

Each task consisted of steps that were scored pass or fail. The score for each task is the number of steps passed. The time required to finish each task was also recorded. Efficiency scores (number of correct steps divided by time) are used in this analysis because of problems associated with the raw score scale [4].

Five administrators tested the automotive mechanics. Two administrators did not administer the hands-on tests for the entire 3-month testing period, so no occasion variable was created (occasion and administrator would be confounded). Efficiency scores were computed for six tasks, treating the alternator-and-battery duty area and the wheel-and-brake duty area as single tasks.

Subjects were nested within administrator, and administrator and task were crossed. The analysis was unbalanced in that the number of subjects was not equivalent for all administrators. The five administrators tested 53, 40, 40, 27, and 13 subjects for a total of 173 test administrations.

The general linear model for the design was:

$$X = A + S(A) + T + AT + TS(A) + e \tag{6}$$

where

$$\begin{aligned} X &= \text{observed score} \\ A &= \text{administrator} \\ S(A) &= \text{subjects within administrator} \\ T &= \text{task} \\ AT &= \text{administrator-by-task interaction} \\ TS(A) &= \text{task-by-subject interaction} \\ e &= \text{error.} \end{aligned}$$

Infantry Rifleman

The hands-on test for the Infantry Rifleman specialty included five duty areas and required about 4 hours to complete. The duty areas, tasks within the duty areas, and point assignments were as follows:

- Target engagement (110 points)
  - Target score
  - Firing upon friendly targets
- First aid (31 points)
  - Stomach wound
  - Jaw wound
  - Arterial bleeding
- Map and compass (85 points)
  - Map
  - Compass
  - Terrain
- Fire team formation (27 points)
  - Symbols
  - Situations
- Antitank mines (53 points)
  - Remove mines
  - Arm mines

The number of points assigned to each duty area varied as shown above and included negative scores for serious errors (e.g., firing upon friendly targets, inability to locate north by reading a compass). A consistent scale of measurement across tasks was created by converting individual scores at the task level to proportion-correct.

Test administrators were not identified on the scoring documents, so no administrator effects can be tested. Likewise, no occasion variable was

created. Accordingly, the general linear model for this specialty is rather simple:

$$X = S + T + TS + e \tag{7}$$

where

X = observed score
S = subject
T = task
TS = task-by-subject interaction
e = error.

ESTIMATION OF EXPECTED MEAN SQUARES AND VARIANCE COMPONENTS

Having identified critical elements of the measurement process that may impact the variance of the performance scores, the task was to determine the composition of the expected mean square for each factor. An expected mean square for a given factor is simply a composite of weighted variance components. The specific variance components included in each mean square are a function of the specification of the general linear model and the status of the effects – either fixed or random. The procedures to determine the appropriate variance-component composition of expected mean squares are discussed in Winer [5].

The composition of the expected mean squares for the random effect models for the three MOSs are given in table 1, table 2, and table 3. The value of any variance-component estimate can be determined by equating the observed mean square to the expected mean square equation and solving for that variance component. These calculations proceed from the bottom of the tables up, because main effects will include variance components involving the interaction components. Such calculations are straightforward

TABLE 1

EXPECTED MEAN SQUARES FOR GROUND RADIO REPAIR TEST

Effect	E(MS)
Occasion [O] p levels	$\sigma_e^2 + \sigma_{BS(OA)}^2 + n\sigma_{OAB}^2 + nq\sigma_{OB}^2 + r\sigma_{S(OA)}^2 + nr\sigma_{OA}^2 + nqr\sigma_O^2$
Administrator [A] q levels	$\sigma_e^2 + \sigma_{BS(OA)}^2 + n\sigma_{OAB}^2 + np\sigma_{AB}^2 + r\sigma_{S(OA)}^2 + nr\sigma_{OA}^2 + npr\sigma_A^2$
OA	$\sigma_e^2 + \sigma_{BS(OA)}^2 + n\sigma_{OAB}^2 + r\sigma_{S(OA)}^2 + nr\sigma_{OA}^2$
Subjects within OA [S(OA)] n subjects	$\sigma_e^2 + \sigma_{BS(OA)}^2 + r\sigma_{S(OA)}^2$
Board [B] r levels	$\sigma_e^2 + \sigma_{BS(OA)}^2 + n\sigma_{OAB}^2 + np\sigma_{AB}^2 + nq\sigma_{OB}^2 + npq\sigma_B^2$
OB	$\sigma_e^2 + \sigma_{BS(OA)}^2 + n\sigma_{OAB}^2 + nq\sigma_{OB}^2$
AB	$\sigma_e^2 + \sigma_{BS(OA)}^2 + n\sigma_{OAB}^2 + np\sigma_{AB}^2$
OAB	$\sigma_e^2 + \sigma_{BS(OA)}^2 + n\sigma_{OAB}^2$
BS(OA)	$\sigma_e^2 + \sigma_{BS(OA)}^2$
Error	σ_e^2

Note: All effects are random.

TABLE 2

EXPECTED MEAN SQUARES FOR
AUTOMOTIVE MECHANIC TEST

Effect	E(MS)
Administrator [A] p levels	$\sigma_e^2 + \sigma_{TS(A)}^2 + n\sigma_{AT}^2 + q\sigma_{S(A)}^2 + nq\sigma_A^2$
Subjects within A [S(A)] n subjects	$\sigma_e^2 + \sigma_{TS(A)}^2 + q\sigma_{S(A)}^2$
Task [T] q levels	$\sigma_e^2 + \sigma_{TS(A)}^2 + n\sigma_{AT}^2 + np\sigma_T^2$
AT	$\sigma_e^2 + \sigma_{TS(A)}^2 + n\sigma_{AT}^2$
TS(A)	$\sigma_e^2 + \sigma_{TS(A)}^2$
Error	σ_e^2

Note: All effects are random.

TABLE 3

EXPECTED MEAN SQUARES FOR
INFANTRY RIFLEMAN TEST

Effect	E(MS)
Subject (S) n subjects	$\sigma_e^2 + \sigma_{ST}^2 + p\sigma_S^2$
Task (T) p levels	$\sigma_e^2 + \sigma_{ST}^2 + n\sigma_T^2$
ST	$\sigma_e^2 + \sigma_{ST}^2$
Error	σ_e^2

Note: All effects are random.

if the research design is balanced. This was not the case for two of the MOSs: Ground Radio Repair and Automotive Mechanic.

Mean squares can still be used to estimate variance components for unbalanced designs, but these methods often involve calculations on the full design matrix, which is a formidable task. The Statistical Analysis System (SAS) computer package contains an efficient procedure for the estimation of variance components of an unbalanced data set [6]. The VARCOMP procedure has been found to produce estimates from simulated unbalanced data that were comparable to the results obtained from a balanced data set [7]. This estimation procedure was used for the two unbalanced data sets of this study.

SECTION 3

RESULTS

VARIANCE-COMPONENT ESTIMATES

Ground Radio Repair

The ANOVA summary and variance-component estimates for the Ground Radio Repair test are presented in table 4. Appendix A provides descriptive statistics for the various combinations of the factors. The magnitude of the variance-component estimate is expressed relative to the total variance. These relative percentages are presented in the last column of table 4.

Over 62 percent of the total variance in the Ground Radio Repair test is accounted for by the residual term $[BS(OA), e]$. This is error variance confounded with variance that is not able to be explained by administrator, occasion, or circuit-board main effects, the interaction of these factors, or individual differences. In other words, a significant percentage of the hands-on test variance *could not* be attributed to the measurement variables that were thought to have impacted the observed variance. Rather, the large residual component indicates that other unidentified sources of error have a large influence on the measurement. Explanations for such a large residual term may include the following:

- Reference materials and technical manuals were available for use by the subjects throughout the testing period. To the extent that the subjects used these materials to guide their troubleshooting on some boards but not on others, the magnitude of the residual term would increase.
- Subjects were encouraged to guess once they had narrowed the possibilities of which component or circuit was faulty. Individuals may

TABLE 4

ANOVA SUMMARY AND VARIANCE-COMPONENT ESTIMATES FOR GROUND RADIO REPAIR TEST

Source of variation	DF	SS	MS		$\hat{\sigma}^2$
Between subjects	88	4401.9	50.0		
Occasion (O)	1	87.3	87.3	0.17	(1.1%)
Administrator (A)	4	367.0	91.7	0.21	(1.4%)
OA	4	181.5	45.4	0.05	(.3%)
Subjects within OA $S(OA)$	79	3766.1	47.7	3.76	(25.1%)
Within subject	801	8562.9	10.7		
Board (B)	9	642.2	71.4	0.46	(3.1%)
OB	9	163.7	18.2	0.24	(1.6%)
AB	36	572.7	15.9	0.37	(2.4%)
OAB	36	378.9	10.5	0.39	(2.6%)
$BS(OA), \epsilon$	711	6805.6	9.6	9.31	(62.3%)

Note: The numbers in parenthesis represent the variance-component estimates expressed as a percentage of the total variance.

have guessed more often on the difficult circuit boards than other boards. Differential guessing by the subjects across the ten circuit boards would have contributed to the large residual term.

- Continuous use of the equipment for a 3-month period may have caused certain grooves or patterns of wear on the circuit boards that would possibly foreshadow the fault to be detected. Some subjects may have been aware of these subtle clues on some boards while other subjects were not.

- To the extent that test administrators were inconsistent in their scoring *within* the same subject (i.e., rated "hard" on some boards and were more lenient on other boards), the residual term would be large.

Other explanations of the large residual terms are possible, particularly explanations that do not involve the measurement factors that were specified in the model for this specialty.

The second largest variance-component estimate was attributable to the subjects within occasion and administrator $[S(OA)]$, which accounted for about 25 percent of the total variance. The circuit-board component (B) accounted for 3.1 percent of the total variance, implying that there were slight mean differences across the ten boards. Given that the circuit boards were from a novel piece of equipment that had not been introduced into the field, it is possible that the subjects had to familiarize themselves with the boards. In effect, the first board served as a practice item and did not necessarily reflect the individual's level of performance on the other nine boards. The other variance-component estimates were negligible, each accounting for less than 3 percent of the variance. Thus, the measurement factors included in the design had little, if any, impact on the variance of the observed hands-on test scores.

Automotive Mechanic

The ANOVA summary and the variance-component estimates for the Automotive Mechanic test are presented in table 5. Efficiency scores were used as the scale of measurement for this hands-on test because the tasks of the hands-on test had differing numbers of steps and varying time limits. Descriptive statistics of the efficiency scores for all tasks and administrators are provided in appendix A.

As with the Ground Radio Repair test, the residual term for this specialty accounted for the majority of the total variance. Many of the same explanations for the magnitude of this term, enumerated earlier, also hold here. However, an occasion factor was *not* included in the measurement model, and, therefore, any variance attributable to this variable is now included in the residual term.

The task factor accounted for 40 percent of the total variance. This large percentage reflects large mean differences in the hands-on efficiency scores across the six tasks (see table A-2 of appendix A). The means ranged from 17.00 for the compression task to 47.98 for the coil task. The magnitude of these mean differences raises questions about their inherent meaning. Are they an artifact of the measurement scale and therefore arbitrary, or are the tasks parallel measures, so that the differences should be considered error? The answer lies in the test-development process – were the tasks developed such that differences in mean scores would have meaning? For this specialty, it is doubtful if the tasks were developed to be parallel measures, particularly given the scale conversion to an efficiency score. A possible solution is to standardize the tasks to have equal means and standard deviations so that the variance due to tasks is zero. However, this has implications for increasing the reliability of the test, as the error variance due to task mean differences is completely removed.

The variance accounted for by individual differences $[S(OA)]$ is excessively small, only 5 percent. The hands-on efficiency scores were not able to "spread" individuals out so that differences in job proficiency could be measured. From other analysis of the hands-on tests [4], it is known that

TABLE 5

ANOVA SUMMARY AND VARIANCE-COMPONENT ESTIMATES FOR AUTOMOTIVE MECHANIC TEST

Source of variation	DF	SS	MS	$\hat{\sigma}^2$	
Between subjects	172	68072.2	395.8		
Administrator (A)	4	7310.5	1827.6	5.2	(1.2%)
Subjects within A $S(A)$	168	60761.7	361.7	22.6	(5.1%)
Within subject	865	359029.0	415.1		
Task (T)	5	156175.4	31235.1	177.2	(40.0%)
AT	20	14219.9	711.0	8.6	(1.9%)
$TS(A), e$	840	188633.7	225.6	229.2	(51.8%)

Note: The measurement scale for these analyses was efficiency score (units correct per hour). The numbers in parenthesis represent the variance-component estimates expressed as a percentage of the total variance.

the Automotive Mechanic test had a ceiling effect; that is, the test was too easy so that many individuals scored extremely well. Certainly, the automotive mechanics were not all equally proficient in performing their job responsibilities. But the test did not have enough difficult items to be able to accurately make these proficiency distinctions.

Infantry Rifleman

Efforts were taken to cluster the tasks of the Infantry Rifleman hands-on test within their respective duty areas (see page 12), but this model could not be estimated because of the tremendous computer memory requirements necessary to solve the problem. The results of the simpler subject-by-task model are given in table 6. Appendix A provides the means for the task factor.

Given such a simple model with only two facets, the residual term can be expected to be large. This residual term included error due to administrator and occasion differences and all other factors that were not specified in the model. The Infantry Rifleman hands-on test included the most diverse range of content (as evidenced by the large task variance-component estimate of almost 18 percent). Given such heterogeneity of test content, it is possible that there was differential performance by subjects across these content areas. If this were the case, the interaction of task and subject (TS) would have been large. However, given the structure of the model, it is not possible to obtain an estimate of this interaction that is independent of the residual variance.

RELIABILITY ESTIMATES

Four reliability estimates were computed for the three hands-on tests: alpha coefficient, ANOVA estimate, and two generalizability coefficients. For each test, the number of items along with the mean and range of the inter-item correlations are noted in table 7. The actual calculations for the estimates presented in appendix B.

TABLE 6

ANOVA SUMMARY AND VARIANCE-COMPONENT ESTIMATES FOR INFANTRY RIFLEMAN TEST

Source of variation	DF	SS	MS	$\hat{\sigma}^2$	
Between subjects					
Subject (S)	258	40.01	0.16	0.008	(10.8%)
Within subject	2849	195.82	0.07		
Task (T)	11	40.02	3.64	0.014	(17.9%)
TS, e	2838	155.80	0.06	0.055	(71.2%)

Note: The variance components were estimated from the mean squares (MS) for each effect. The numbers in parenthesis represent the variance-component estimate expressed as a percentage of the total variance.

TABLE 7

ALTERNATIVE ESTIMATES OF RELIABILITY
FOR THREE HANDS-ON TESTS

	Ground Radio Repair	Automotive Mechanic	Infantry Rifleman
Number of items	10 boards	6 tasks	12 tasks
Inter-item correlation			
Mean	0.28	0.12	0.14
Range	0.00 to 0.59	0.01 to 0.27	-0.15 to 0.64
Alpha estimate	0.80	0.40	0.65
ANOVA estimate	0.79	- -[a]	0.56
G-theory estimate			
Relative	0.80	0.37	0.64
Absolute	0.77	0.25	0.58

a. The ANOVA reliability estimate for the Automotive Mechanic specialty could not be determined because the within-subject mean square was greater than the between-subjects mean square. See table 5.

The reliability estimates for the Ground Radio Repair specialty were in the moderately high range, 0.77 to 0.80. Such results would be expected given the consistency of the test content – troubleshooting ten circuit boards. Essentially, the test is composed of ten parallel measures or replications. The reliabilities were not higher due to the large residual variance-component estimate. No differences were noted among the four separate reliability estimates for this specialty. This is because of the insignificant contribution of each measurement factor to the explanation of the total variance.

Unacceptable reliabilities were noted for the Automotive Mechanic hands-on test. Given that all estimates were less than 0.50, more variance in the observed scores was accounted for by random error than by true score differences. In fact, for the ANOVA estimate, the within-subjects mean square was greater than the between-subjects mean square. This contrary finding was due to the large contribution of the task variance-component estimate to the within-subjects variance. Likewise, the negligible percentage of individual-difference variance, due to the ceiling effect on the score scale, restricted the overall magnitude of the reliabilities.

The reliability results for the Infantry Rifleman specialty were in the low range – that is, 0.56 to 0.65. This reflects the diversity of test content for this hands-on test; inter-item correlations ranged from -0.15 to 0.64. The ANOVA estimate for this specialty was the lowest of the four estimates because of the large magnitude of the task main effect and residual term (both are within-subject factors). The relative generalizability coefficient was the same as the alpha coefficient because it did not include the variance due to the task facet.

SECTION 4

DISCUSSION

The residual variance-component estimates for the three hands-on tests were large. This is unfortunate because this is the term that should be kept to a minimum to enhance the reliability of the hands-on tests. The implication of such large estimates is that the models were not correctly specified and therefore did not include the appropriate measurement factors to account for the significant residual variances.

APPROPRIATE EXPERIMENTAL DESIGNS FOR RELIABILITY ASSESSMENT

In the case of the three specialties of this study, the issue of large residual variances may not be so much the misspecification of the model as the inappropriateness of the data-collection effort. As stated earlier, the data collection was not designed for these particular G-theory analyses. Therefore, the data tended to be unbalanced and nested. Given that the purpose of the G-theory analysis is to be able to estimate as many variance components as possible, the 1981 data-collection designs did not allow for the estimation of many important variance components. An example will illustrate this point.

For two of the specialties, the subjects factor is nested within administrator, that is, each subject is tested once and by only one administrator. In this respect, the mean hands-on score for any administrator is confounded with the ability level of the group of subjects that he tests. Random assignment of subjects generally assumes that the ability differences among the groups of subjects tested by any administrator is insignificant; however, this may not always be the case. Also, computational problems arise when the subjects are unbalanced with respect to their nesting within administrator; that is, differing numbers of subjects are tested by each administrator. This was particularly the case for the Automotive Mechanic specialty: one

administrator tested 53 subjects while another administrator tested only 13. Likewise, by having subjects nested within administrator, the subject-by-administrator-interaction variance component cannot be estimated. A better research design would not involve this nesting of subjects. Rather, it would be more efficient if the multiple administrators tested each subject. In this manner, a pure estimate of both the subject and administrator factors and their interaction could be determined.

In summary, it is best that the number of between-subjects factors be kept to a minimum. This will reduce the nesting of subjects and allow for the proper estimation of the variance component for this term. If there is any nesting of factors, either for the between or within factors, it is necessary that the nested levels be balanced. Likewise, multiple administrators should rate each subject's performance on the hands-on tasks.

SCALING IMPLICATIONS FOR RELIABILITY

The score scales of the three hands-on tests are dramatically different:

- Ground Radio Repair: sum of three component scores. The total score for each circuit board ranges from 0 to 14.
- Automotive Mechanic: efficiency score. Tasks have significantly different means and standard deviations.
- Infantry Rifleman: proportion-correct score. The correct number of steps are divided by the total number of scorable steps.

In the classical definition of reliability, task scores are considered to be parallel measures. In this regard, mean differences across the tasks as well as differences in the rank ordering of individuals by the tasks (task-by-subject interaction) are considered error. Given the above score scales for each of the hands-on tests, it is questionable whether the tests were developed so that the tasks would have equivalent means. Therefore, mean differences

across the tasks of these tests are an artifact of the test-construction process, and thus are arbitrary and have no inherent meaning.

The best example of these arbitrary scaling differences is the Automotive Mechanic hands-on test. As noted in table A-2, appendix A, the task means and even standard deviations are very different. Accordingly, the task variance-component estimate for this test was exceptionally large, accounting for 40 percent of the total variance. As mentioned earlier, an alternative that would correct the scale arbitrariness is to standardize each task to a common mean and standard deviation. The effect of task standardization is to delete the variance due to task differences (i.e., $\hat{\sigma}^2_{task} = 0$) and increase the reliability of the test because task differences no longer contribute to the error component. In fact, the reliability estimates increase due to standardization as follows: alpha, 0.40 to 0.48; relative estimate, 0.37 to 0.45; and absolute estimate, 0.25 to 0.44. These reliabilities, although still substandard, did increase a relatively large amount.

In summary, the reliability of an instrument should be "built in," not be an afterthought. Therefore, reliability should be addressed as part of the test-development process. In particular, the construction (or lack of construction) of a particular score scale for the tasks has implications for the reliability. If at all possible, it is preferable that these scales have comparable means and standard deviations, or if the scales do differ, that the differences do have meaning and are not arbitrary. This is a task not easily accomplished, but one frequently left to chance, as was the case in the development of the hands-on tests of this study.

IMPLICATIONS FOR THE MARINE CORPS JOB PERFORMANCE MEASUREMENT PROJECT

The use of experimental designs to assess reliability (i.e., G-theory analysis) is extremely difficult to coordinate, expensive and time consuming to administer, and often disruptive to personnel. In recognition of these constraints, it is recommended that small G-studies be conducted on limited samples – that is, 20 to 30 subjects. Within this context, G-theory studies

are proposed to address research questions pertinent to the full-scale Marine Corps testing of the Infantry occupational field.

Although not a specific finding in this study, the inconsistency among the test administrators is considered to be a major threat to the fidelity of hands-on measurement of job performance. Therefore, explicit examination of the administrators' scoring strategies and practices are the central focus of the G-theory experimental designs. Three specific G-theory studies are discussed to address this concern of administrator scoring consistency. Although the three study designs are not mutually exclusive and certainly can be changed to accomodate the requirements of the Marine Corps, they are presented as three independent efforts.

Consistency Between Administrators

Consistency of scoring hands-on performance by different administrators is a primary concern. An individual's test score should not vary depending on which administrator rated his performance. Administrators should be like scoring keys used for paper-and-pencil tests – uniformly and equably grading the quality of performance. The first experimental design, presented in figure 1, addresses this question of inter-administrator reliability. In addition, the design examines the extent to which alternate forms of the hands-on test are parallel and whether there is an order effect of taking one form of the test before the other.

Note from figure 1 that N subjects are administered two test forms (A and B). Each test form is composed of ten different tasks; thus, tasks are nested within test form and crossed with subjects. The test forms are administered in counter-balanced order so that half of the subjects first take form A and then form B, and the other half of the sample take B and then A. In this manner, subjects are nested within order. The same three administrators rate each subject for both test forms, so that administrators are crossed with subjects and tasks. More than three administrators would be benefical for statistical purposes, but the number of administrators should be limited to a feasible number of persons who can independently observe

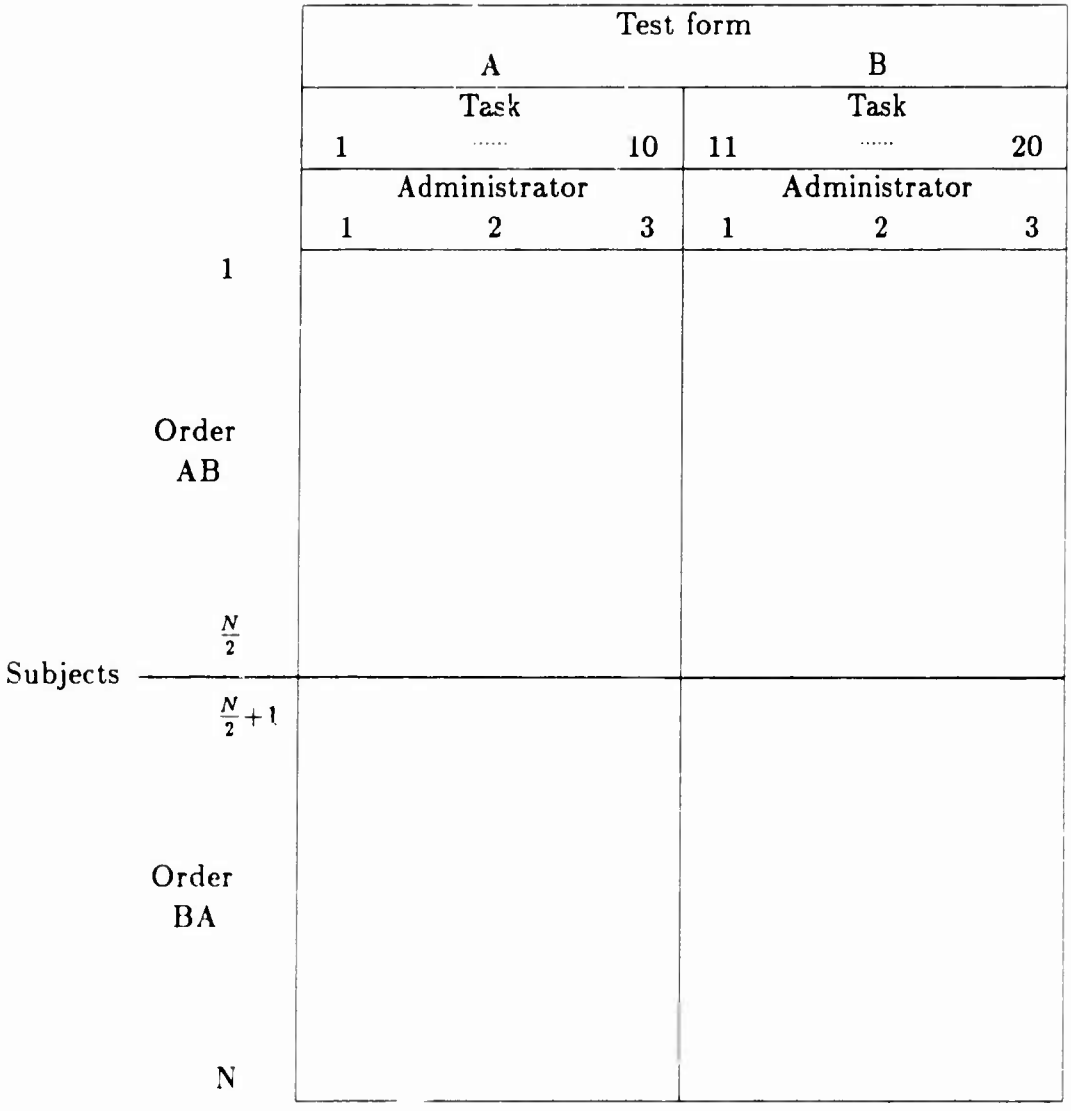

FIG. 1: EXPERIMENTAL DESIGN FOR CONSISTENCY BETWEEN ADMINISTRATORS

one individual perform the tasks. This experimental design is completely balanced in that there is an equal number of tasks within each test form and an equal number of subjects within each testing order.

Given the above design description, the following facets are able to be estimated. Not all interaction terms are of interest and therefore are included in the residual term.

Between subjects		Within subject	
Order	$[O]$	Form	$[F]$
Subjects(order)	$[S(O)]$	Task(form)	$[T(F)]$
		Administrator	$[A]$
		Form-by-administrator	$[FA]$
		Task(form)-by-administrator	$[T(F)A]$
		Order-by-form	$[OF]$
		Order-by-administrator	$[OA]$
		Subjects(order)-by-form	$[S(O)F]$
		Subjects(order)-by-administrator	$[S(O)A]$
		Residual	$[e]$

With respect to the question of inter-administrator reliability, the variance-component estimates have the following interpretation:

A	mean differences among administrators
FA	administrator inconsistencies across test forms
$T(F)A$	administrator inconsistencies across tasks within test form
OA	administrator inconsistencies for the testing order
$S(O)A$	administrator inconsistencies between and within subject

The issue of parallel forms is addressed primarily by the variance-component estimate for forms (F) and other analyses to be conducted independent of these studies. In addition, the other estimates involving the forms facet (FA, OF, and $S(O)F$) provide some information about form equivalence.

It should be noted, however, that tasks do not necessarily have to be equivalent within forms. In other words, the task-within-form component $(T(F))$ can be large.

Two assumptions are made concerning the application of this design. First, subjects are randomly assigned to a testing order. Since subjects are nested within order, it is necessary that differences attributed to the order effect are not a function of the specific individuals who were tested within each order. Random assignment of subjects should dissipate this potential influence. Second, it is necessary that no time or only a short time interval separate the administration of the two forms. The question of form equivalence should not be confounded with testing occasion or stability of the performance construct over time.

Administrator Consistency Over Time

Test administrators are trained extensively at the beginning of the project to be consistent scorers of performance. As they begin testing subjects, they gain experience and encounter situations they may not have been exposed to in training. Each administrator develops his own scoring strategies that possibly vary depending on the exact time within the testing period. Therefore, it is necessary to determine the extent to which administrators drift over time and whether this drift contributes variance to an individual's hands-on score. In addition, it is necessary to obtain an estimate of how well an individual would do if the test had been administered at a different time. These questions of administrator drift over time and test-retest reliability can be simultaneously examined within the context of a G-theory design.

Figure 2 provides the data layout for examining differences among three test administrators across two different testing occasions. In this design, each subject takes only one test form but is retested with the same form on a second occasion. The time interval between these two test administrations should not exceed 2 weeks to eliminate the possibility of relevant, intervening job experiences. An equal number of subjects takes each form.

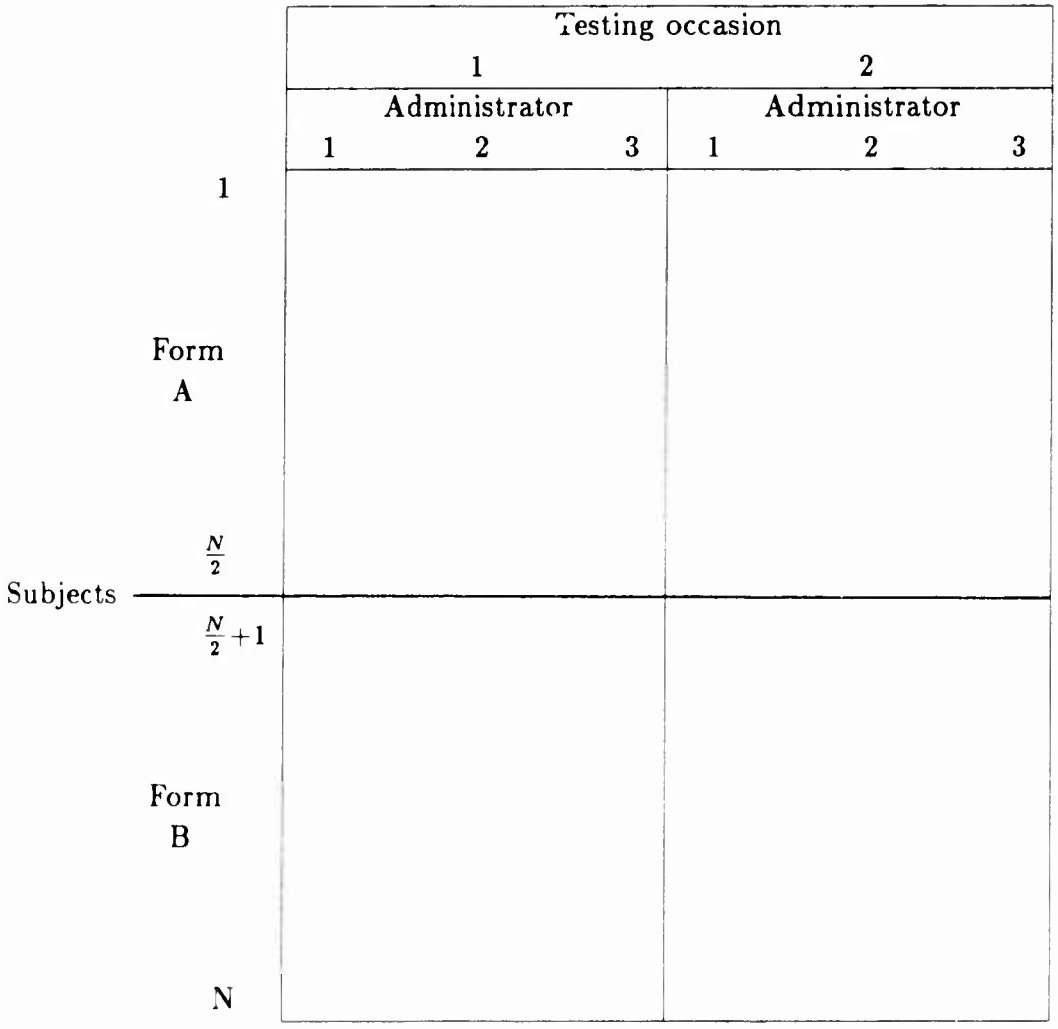

FIG. 2: EXPERIMENTAL DESIGN FOR ADMINISTRATOR CONSISTENCY OVER TIME

Therefore, the design is balanced with subjects nested within form. Administrators are crossed with testing occasion and both factors are crossed with subjects and form. Again, it is necessary that subjects are randomly assigned to the form they will be administered.

The design shown in figure 2 will result in the following variance-component estimates:

Between subjects		Within subject	
Form	$[F]$	Occasion	$[O]$
Subjects(form)	$[S(F)]$	Administrator	$[A]$
		Occasion-by-administrator	$[OA]$
		Form-by-occasion	$[FO]$
		Form-by-administrator	$[FA]$
		Form-by-occasion-by-administrator	$[FOA]$
		Subjects(form)-by-occasion	$[S(F)O]$
		Subjects(form)-by-administrator	$[S(F)A]$
		Residual	$[S(F)OA, e]$

This design includes all effects that can be estimated. Therefore, the residual term is confounded with the three-way interaction of subjects within form-by-occasion and administrator.

If the occasion and form-by-occasion components are small relative to the total variance, then test-retest reliability will be high. Also, if the occasion-by-administrator component is small, then administrators will tend to be consistent across the time interval of 2 weeks. Other studies could be designed to address the issue of administrator drift over a longer period. In addition, much of this design replicates the first study in that administrator inconsistencies are examined again (A, OA, FA, and $S(F)A$).

Administrator Consistency Across Bases

The Infantry occupational field will be the focus of the Marine Corps' initial effort in the Job Performance Measurement Project. Testing of the infantry will be conducted at two bases: Camp Lejeune, NC, and Camp Pendleton, CA. If test administrators are not using the same scoring standards at these two bases, then essentially two different concepts of job performance are being measured. An infantryman's score on the hands-on test, in this case, would partially depend on the base at which he is stationed.

To explicitly address this question of scoring differences across bases, the G-theory design in figure 3 is proposed. The design requires that three administrators travel from their home base to be cross-trained at the other base and then participate in the tryout of the hands-on test at this opposite base. For example, three administrators from Camp Lejeune would travel to Camp Pendleton to be trained. During the tryout, they would score the performance of subjects along with three administrators from Camp Pendleton. The same process would occur for a group of three Camp Pendleton administrators traveling to Camp Lejeune. Although this design does not fully cross the administrators and subjects (that is, the same six administrators do not test all N subjects), it does eliminate the problems associated with training the test administrators twice. Therefore, there is no question of order effects. That is, the base at which the training and tryout occurred first would certainly be different from the base where this occurred second. Other designs can be implemented that will impose the dual training of the administrators. Given the burden (as well as boredom) of additional training and testing placed on the administrators, this approach is not advocated.

Because administrators are not completely crossed with subjects, the analysis can be done within base, so that the analysis of the second base is a replication. The variance components that can be estimated from the design of figure 3 are the following:

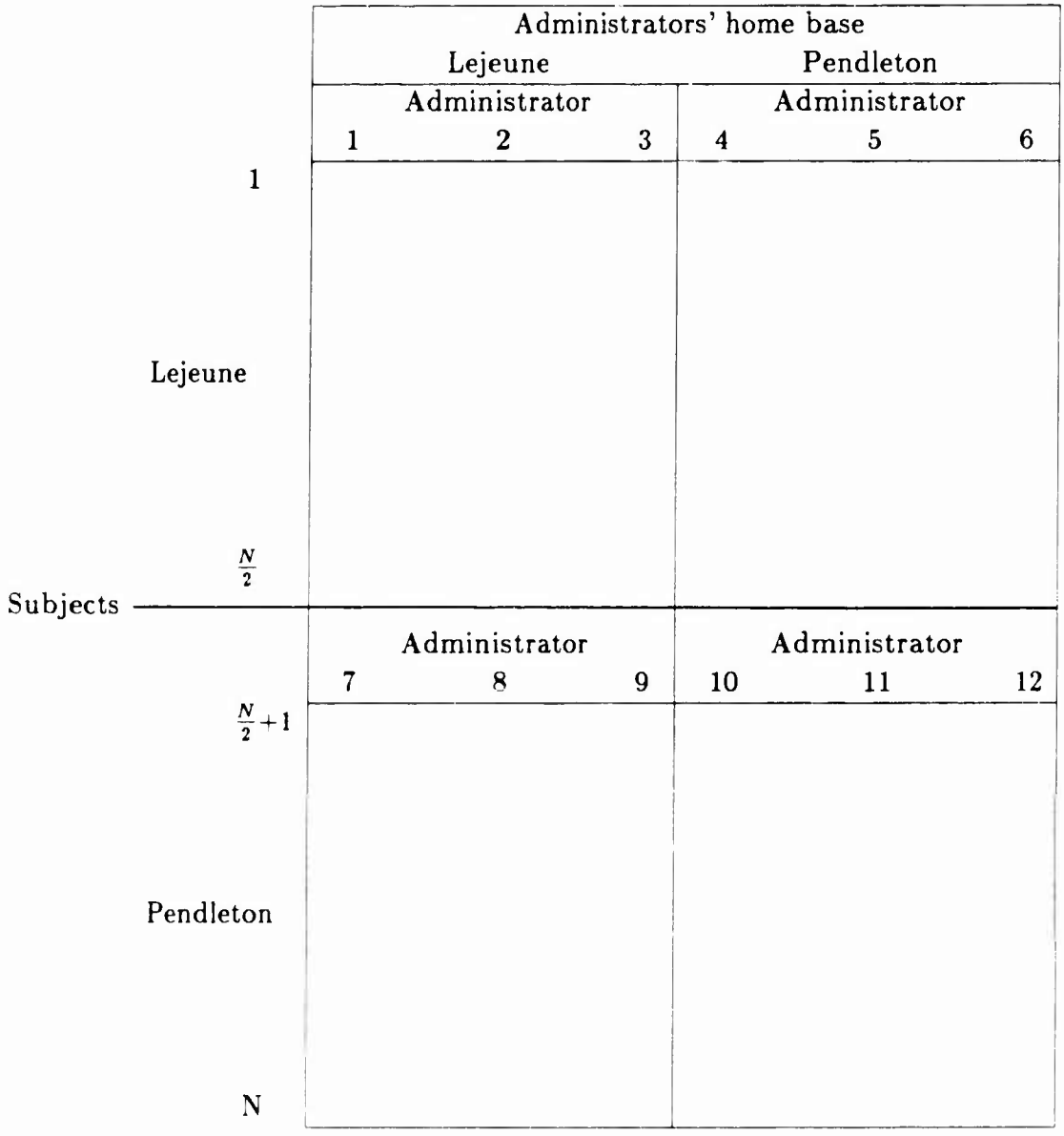

FIG. 3: EXPERIMENTAL DESIGN FOR ADMINISTRATOR CONSISTENCY ACROSS BASES

Between subjects		Within subject	
Subjects	$[S]$	Home base	$[H]$
		Administrator(home)	$[A(H)]$
		Subjects-by-home	$[SH]$
		Residual	$[SA(H), e]$

The variance components of most interest are the home-base main effect (H) and the interaction of this term with subjects (SH). Both terms will flag any discrepancies among the administrators as they score performance. The home-base effect will note if the two teams of three administrators are different. Also, the magnitude of the interaction effect will show if administrators score their "own" subjects differently from the subjects of the other base.

CONCLUSIONS

- Generalizing from performance on the hands-on tests to performance in the three MOSs is limited, given the magnitude of the residual variance in the hands-on scores.
- Experimental designs that address the impact of specific measurement factors on the variance of the hands-on scores should be developed. Such designs allow for the simultaneous consideration of several factors and their interactions, as well as the calculation of generalizability coefficients.

REFERENCES

[1] Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnum, N. *The Dependability of Behavioral Measurements.* New York: Wiley, 1972

[2] Brennan, R. L. *Elements of Generalizability Theory.* Iowa City: ACT Publications, 1983

[3] CNA, Report 89, *An Evaluation of Using Job Performance Tests to Validate ASVAB Qualification Standards*, by Milton H. Maier and Catherine M. Hiatt, Unclassified, May 1984

[4] CNA, Research Contribution 540, *Examining the Validity of Hands-On Tests of Job Performance*, by Paul W. Mayberry, Unclassified, May 1986

[5] Winer, B. J. *Statistical Principles In Experimental Design.* New York: McGraw-Hill, 1962

[6] SAS Institute Inc. *SAS User's Guide: Statistics, Version 5 Edition.* Cary, NC: SAS Institute Inc., 1985

[7] Bell, J. F. "Generalizability Theory: The Software Problem," *Journal of Educational Statistics* 10 (Spring 1980): 19-29

[8] Kirk, R. E. *Experimental Design: Procedures for the Behavioral Sciences.* Belmont, CA: Brooks/Cole Publishing Co., 1968

APPENDIX A

DESCRIPTIVE STATISTICS FOR HANDS-ON TESTS

TABLE A-1

DESCRIPTIVE STATISTICS FOR THE
GROUND RADIO REPAIR TEST

Panel A:
Administrator by Testing Occasion

Administrator	Testing occasion						Total		
	1			2					
	N	Mean	SD	N	Mean	SD	N	Mean	SD
1	7	11.86	1.93	9	12.07	2.88	16	11.97	2.44
2	5	13.52	0.78	8	12.62	1.71	13	12.97	1.45
3	10	11.54	2.01	9	12.39	1.83	19	11.94	1.92
4	10	11.49	2.00	11	11.85	2.30	21	11.68	2.11
5	12	9.90	2.82	8	11.95	1.91	20	10.72	2.65
Total	44	11.36	2.34	45	12.16	2.11	89	11.76	2.25

Panel B:
Circuit Board by Testing Occasion

Board	Testing occasion				Total	
	1		2			
	Mean	SD	Mean	SD	Mean	SD
1	13.32	1.83	12.89	2.58	13.10	2.24
2	12.59	3.13	12.44	3.62	12.52	3.36
3	10.68	3.88	11.47	3.42	11.08	3.66
4	11.59	4.26	12.27	3.95	11.93	4.09
5	12.50	2.93	12.80	2.57	12.65	2.74
6	10.55	3.64	10.91	3.59	10.73	3.60
7	11.32	4.04	12.22	3.70	11.78	3.88
8	8.86	5.22	12.04	3.75	10.47	4.79
9	12.09	4.25	12.53	3.55	12.31	3.90
10	10.07	5.14	11.98	3.84	11.03	4.61
Total	11.36	4.12	12.16	3.50	11.76	3.84

TABLE A-1 (continued)

Panel C:
Circuit Board by Administrator

Board	Test administrator									
	1		2		3		4		5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	11.87	3.22	13.23	2.77	13.79	0.63	13.14	1.96	13.30	1.98
2	12.00	4.38	12.46	4.10	13.79	0.92	12.48	2.60	11.80	4.05
3	10.75	4.12	10.92	3.71	11.16	2.93	11.71	3.42	10.70	4.32
4	11.75	4.84	13.08	1.93	12.74	2.51	12.86	3.32	9.60	5.57
5	12.25	3.26	14.00	0.00	12.84	2.43	12.19	3.09	12.40	3.02
6	10.87	3.50	12.15	3.11	10.42	4.19	9.76	3.10	11.00	3.87
7	13.75	0.68	13.69	1.11	10.63	4.62	11.14	4.22	10.70	4.55
8	12.12	4.22	12.77	3.42	10.32	4.53	10.00	4.94	8.30	5.32
9	12.37	4.46	13.69	1.11	12.63	4.11	12.00	3.35	11.40	4.86
10	12.00	3.93	13.69	1.11	11.11	4.28	11.48	4.12	8.00	5.88
Total	11.98	3.80	12.97	2.65	11.94	3.58	11.68	3.59	10.72	4.67

TABLE A-2

DESCRIPTIVE STATISTICS FOR THE
AUTOMOTIVE MECHANIC TEST

Administrator by Task Efficiency Scores

Administrator	N	Compression		Coil		Vacuum	
		Mean	SD	Mean	SD	Mean	SD
1	53	16.25	6.16	43.72	16.56	28.78	14.58
2	40	19.97	6.51	46.50	15.04	24.88	9.35
3	40	16.04	4.90	46.59	17.35	18.89	8.67
4	13	16.07	5.56	54.66	16.72	24.98	13.79
5	27	15.96	6.50	57.38	23.83	29.81	17.00
Total	173	17.00	6.14	47.98	18.21	25.47	13.21

Administrator	Precision timing		Alternator and battery		Wheel and brake		Total	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	41.31	12.37	28.74	18.68	54.97	26.18	35.63	20.94
2	42.23	11.94	21.06	6.76	46.80	28.08	33.57	18.90
3	44.40	12.59	20.81	9.28	38.11	21.49	30.81	18.42
4	47.45	32.02	22.33	9.02	32.35	12.11	32.97	21.60
5	48.06	15.23	30.72	9.16	51.99	22.01	38.99	22.09
Total	43.75	15.09	24.96	13.09	47.02	25.14	34.36	20.29

TABLE A-3

DESCRIPTIVE STATISTICS FOR THE
INFANTRY RIFLEMAN TEST

Task	Mean	SD
Target score	.55	.23
Firing upon friendly targets	.16	.20
Stomach wound	.51	.21
Jaw wound	.50	.34
Arterial bleeding	.59	.18
Map	.50	.23
Compass	.55	.38
Terrain	.42	.18
Symbols	.59	.22
Situations	.52	.25
Remove mines	.44	.24
Arm mines	.58	.27
Total	.49	.28

Note: N = 259 infantry riflemen. The score scale is proportion correct.

APPENDIX B

CALCULATION OF RELIABILITY ESTIMATES FOR THREE MARINE CORPS MOSs

APPENDIX B

CALCULATION OF RELIABILITY ESTIMATES FOR THREE MARINE CORPS MOSs

Four estimates of reliability were computed for each of the three Marine Corps MOSs:

- Alpha coefficient
- ANOVA estimate
- G-theory relative generalizability coefficient
- G-theory absolute generalizability coefficient.

The data for these calucalations are noted throughout the report but primarily are from appendix A, tables A-1 to A-3, and the ANOVA-variance component summaries, tables 4 through 6.

Alpha Estimate

The general formula for the alpha coefficient is

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum S^2_{X_i}}{S^2_{Total}}\right] \qquad (B-1)$$

where

S^2_{Total} = the observed score variance for a test composed of n components that each have a variance $S^2_{X_i}$.

The observed score variances for the total score were not presented in the text or appendix A. These variances are as follows: Ground Radio Repair, 506.42; Automotive Mechanic, 2374.61; and Infantry Rifleman, 1.86.

Ground Radio Repair

$$\alpha = \frac{10}{10-1}\left[1 - \frac{5.0 + 11.3 + 13.4 + 16.7 + 7.5 + 13.0 + 15.1 + 22.9 + 15.2 + 21.3}{506.4}\right] = 0.80$$

Automotive Mechanic

$$\alpha = \frac{6}{6-1}\left[1 - \frac{37.7 + 331.6 + 174.5 + 227.7 + 171.3 + 632.0}{2374.6}\right] = 0.40$$

Infantry Rifleman

$$\alpha = \frac{12}{12-1}\left[1 - \frac{.05 + .04 + .04 + .12 + .03 + .05 + .14 + .03 + .05 + .06 + .06 + .07}{1.9}\right] = 0.$$

ANOVA Estimate

The general formula for the ANOVA estimate of reliability is

$$\rho_{XX'} = 1 - \frac{MS\ within\ subject}{MS\ between\ subjects} \quad . \qquad (B-2)$$

Ground Radio Repair

$$\rho_{XX'} = 1 - \frac{10.7}{50.0} = 0.79.$$

Automotive Mechanic

A reliability estimate cannot be computed by the ANOVA procedure because MS within subject (415.1) is greater than MS between subjects (395.8).

Infantry Rifleman

$$\rho_{XX'} = 1 - \frac{0.07}{0.16} = 0.56.$$

G-Theory Relative Estimate

The general formula for the G-theory relative generalizability coefficient is

$$\rho^2(Rel) = \frac{\sigma^2_{Subject}}{\sigma^2_{Subject} + \Sigma\sigma^2_{Subject\ interaction\ terms/n'}} \qquad (B-3)$$

where

n' = the respective sampling frequency of each source of error.

Ground Radio Repair

$$\rho^2(Rel) = \frac{3.76}{3.76 + 9.31/10} = 0.80$$

Automotive Mechanic

$$\rho^2(Rel) = \frac{22.6}{22.6 + 229.2/6} = 0.37$$

Infantry Rifleman

$$\rho^2(Rel) = \frac{.008}{.008 + .055/12} = 0.64$$

G-Theory Absolute Estimate

The general formula for the G-theory absolute estimate of reliability is

$$\rho^2(Abs) = \frac{\sigma^2_{Subject}}{\sigma^2_{Subject} + \Sigma\sigma^2_{Main\ effects\ and\ interaction\ terms/n'}} \qquad (B-4)$$

where

n' = the respective sampling frequency of each source of error.

Ground Radio Repair

$$\rho^2(Abs) = \frac{3.76}{3.76 + \frac{.17}{2} + \frac{.21}{5} + \frac{.05}{10} + \frac{.46}{10} + \frac{.24}{20} + \frac{.37}{50} + \frac{.39}{100} + \frac{9.31}{10}} = 0.77$$

Automotive Mechanic

$$\rho^2(Abs) = \frac{22.6}{22.6 + \frac{5.2}{5} + \frac{177.2}{6} + \frac{8.6}{30} + \frac{229.2}{6}} = 0.25$$

Infantry Rifleman

$$\rho^2(Abs) = \frac{.008}{.008 + \frac{.014}{12} + \frac{.055}{12}} = 0.58$$